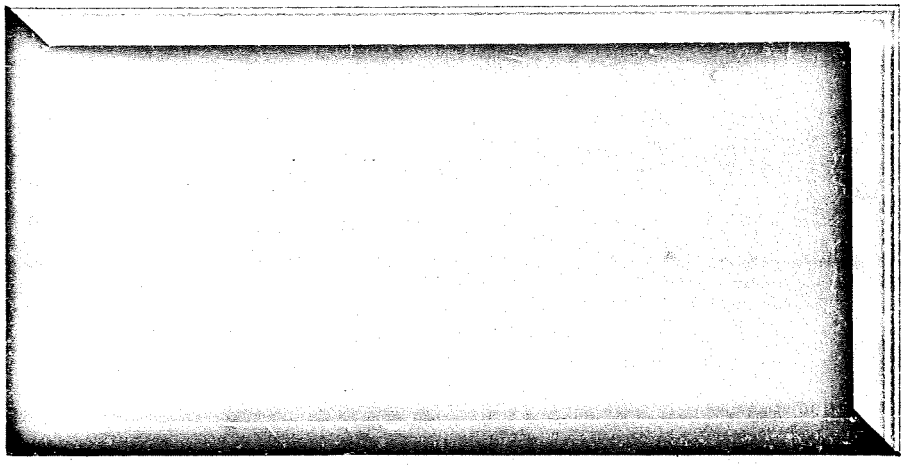

(NASA-CR-151321) INTEGRATED MEDICAL AND BEHAVIORAL LABORATORY MEASUREMENT SYSTEM, PHASE B 2. VOLUME 3: SYSTEM CONCEPT AND DESIGN. APPENDIX E: PRECISION AND ACCURACY REPORT (General Electric Co.) 10 p

N77-78303

Unclas

00/12 29913

DOCUMENT NO. 68SD8014
12 FEBRUARY 1968

INTEGRATED MEDICAL AND BEHAVIORAL LABORATORY MEASUREMENT SYSTEM PHASE B II FINAL REPORT

VOLUME III - SYSTEM CONCEPT AND DESIGN

APPENDIX E

PRECISION AND ACCURACY REPORT (REPRINT) *

CONTRACT NASW—1630

PREPARED FOR THE

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

BY THE

BIOASTRONAUTIC SECTION

SPACE SYSTEMS ORGANIZATION

GENERAL ELECTRIC

MISSILE AND SPACE DIVISION
Valley Forge Space Technology Center
P.O. Box 8555 • Philadelphia 1, Penna.

*REPRINTED BY PERMISSION OF THE PUBLISHERS, THE WILLIAMS AND WILKINS CO., FROM THE AMERICAN JOURNAL OF CLININAL PATHOLOGY, VOLUME 43, NO. 6, PAGES 562-569, 1965.

THE AMERICAN JOURNAL OF CLINICAL PATHOLOGY
Copyright © 1965 by The Williams & Wilkins Co.

Vol. 43, No. 6
Printed in U.S.A.

A SCHEME FOR THE COMPARISON OF QUANTITATIVE METHODS

ROY N. BARNETT, M.D.

Division of Laboratories, Norwalk Hospital, Norwalk, Connecticut

The scheme described below was developed as part of the program of the Standards Committee of the College of American Pathologists in response to the need for a statistical method by which new diagnostic products could be evaluated. It was found to be so satisfactory for this purpose that it is presented here in a form which permits a more general application. Although the discussion and examples are chosen to reflect a clinical laboratory orientation, the scheme is equally applicable to many nonclinical types of quantitative analyses.*

In brief the scheme uses a number of simple statistical tools by means of which a quantitative method of analysis (test method) can be evaluated in a single laboratory by comparison with a reference method. The reference method may be the one currently employed in the laboratory or a different one thought to be more suitable for some reason.

APPLICABILITY

The scheme is applicable to most substances present in blood or body fluids. The material for which the analysis is performed must be such that it can be prepared in a reasonably pure form and can be defined quite specifically. For example, glucose, urea, cholesterol, and calcium are suitable. Enzyme methods can be compared satisfactorily only if pH of the reaction mixture, substrate, direction of reaction, expression of units, etc., are identical, and this is rarely the case. If the methodology itself causes different substances to be given the same name, as for example, protein fractions isolated by different technics, the scheme may not be applicable; on the other hand, use of the scheme may be useful in defining the problem precisely.

Received, September 12, 1964.

* The statistical consultant was Dr. W. J. Youden of the National Bureau of Standards.

COLLECTION OF DATA

Four types of data are required and for each of these the appropriate form is shown. Figures for a hypothetical series of analyses have been entered in order to illustrate the arithmetic.

A. *Reproducibility form* (*Table 1*). This involves analyses of purified material in solution at 3 different levels: low, intermediate, and high. The levels are chosen arbitrarily for convenience and relation to the usual levels found in patients. In the example shown glucose levels were chosen at 50 mg., 100 mg., and 200 mg. per 100 ml. For chlorides, 70, 100, and 130 mEq. per liter would be more suitable. Each method, the test and the reference, is entered on its own form. The standards are prepared, and on each of 10 separate days each standard is analyzed by each method. Although one can "speed up" the procedure by performing the 10 determinations in less than 10 working days, one would thereby lose some of the variability ordinarily introduced into analyses by changes in technologist performance, weather, reagents, instruments, etc. Such a "speed-up" is not recommended.

B. *Recovery experiment form* (*Table 2*). Recovery experiments are performed at 3 different levels. The material to be tested, for example, serum or urine, is used as a base. To aliquots of this are added enough pure material to produce final dilutions elevated by approximately 20, 50, and 100 per cent of the usual normal value. It is imperative that this be done in such a manner that the added material actually goes into solution. In the example chosen we assumed a normal blood glucose of about 80 mg. per 100 ml. The final concentrations then were original (76 mg. by test) plus 16 mg.; original plus 40 mg.; and original plus 80 mg. Each determination is made in *triplicate* by means of each method. Results are charted on separate forms and calculations performed as indicated thereon.

C. *Data form* (*Table 3*). Routine samples

TABLE 1

REPRODUCIBILITY FORM

Replicate Determination	Low (50 mg.)	Date	Intermediate (100 mg.)	Date	High (200 mg.)	Date
1	48	June 4	88	June 4	220	June 4
2	54	June 5	90	June 5	204	June 5
3	52	June 6	106	June 6	188	June 6
4	41	June 7	104	June 7	197	June 7
5	47	June 8	98	June 8	194	June 8
6	56	June 11	104	June 11	206	June 11
7	52	June 12	92	June 12	194	June 12
8	50	June 13	100	June 13	216	June 13
9	47	June 14	94	June 14	201	June 14
10	53	June 15	97	June 15	191	June 15
Average ($\bar{x}$)	50.0		97.3		201.1	
S.D. $= \sqrt{\frac{S(x - \bar{x})^2}{N - 1}}$	4.37		6.25		10.56	
Range	±15 per cent		±9.2 per cent		±8 per cent	

S.D.: This is the usual formula. Subtract the mean ($\bar{x}$) from each value (x). Square each of these differences. Add the squares. Divide by 9. Take the square root of this value.

Range: Subtract lowest from highest value in each group of 10. Divide by 2. Multiply by $\frac{100}{\bar{x}}$. Express as ± the per cent so obtained.

Analyst: *J. Smith M.T. (ASCP)*
Name of Method: *Glucose-CD*

TABLE 2

RECOVERY EXPERIMENT FORM*

	Column 1 Value Obtained by Test	Column 2 Observed Difference from Value A	Column 3 Expected Difference from Value A	Column 4 Per Cent Recovery†
A. Original sample	76	0	0	None
B. Same plus approximately 20 per cent of normal	90	14	16	87
C. Same plus approximately 50 per cent of normal	117	41	40	102
D. Same plus approximately 100 per cent of normal	154	78	80	97

* Each determination is done in triplicate and the results averaged.
† Column 2 value divided by column 3 value times 100.

Analyst: *J. Smith, M.T. (ASCP)*
Method Used for Testing: *Glucose-CD*
Date: *June 14*

from the proper population, for example hospital patients, are used. Forty samples are analyzed with both methods and the required data entered in columns 1 through 5. Not more than 5 tests per day should be performed. During this phase of the study obvious "large discrepancies" between the 2 methods may be encountered. Such values are entered on a large discrepancy form (Table 4; note example; also see LD point on Figure 1) and are *not* included in the 40 replicate determinations. Every effort is

TABLE 3
DATA FORM

1 No. of Test	2 Sex of Patient	3 Date	4 Result Reference Method	5 Result Test Method	6 Difference 4–5	7 Difference Minus Bias	8 Column 7 Squared
1	M	June 18	180	195	−15	−11.1	123.2
2	M	June 18	145	142	+3	+6.9	47.6
3	F	June 18	47	54	−7	−3.1	9.6
4	M	June 18	83	83	0	+3.9	15.2
5	F	June 19	91	91	0	+3.9	15.2
6	F	June 19	70	78	−8	−4.1	16.8
7	F	June 20	206	217	−11	−7.1	50.4
8	M	June 20	63	74	−11	−7.1	50.4
9	F	June 20	67	63	+4	+7.9	62.4
10	F	June 20	94	92	+2	+5.9	34.8
11	M	June 21	108	112	−4	−0.1	0
12	M	June 22	61	67	−6	−2.1	4.4
13	F	June 22	334	345	−11	−7.1	50.4
14	F	June 22	144	142	+2	+5.9	34.8
15	M	June 22	71	75	−4	−0.1	0
16	F	June 25	110	110	0	+3.9	15.2
17	F	June 25	134	144	−10	−6.1	37.2
18	F	June 25	101	102	−1	+2.9	8.4
19	M	June 25	69	70	−1	+2.9	8.4
20	M	June 25	184	197	−13	−9.1	82.8
21	F	June 26	78	78	0	+3.9	15.2
22	M	June 26	54	65	−11	−7.1	50.4
23	M	June 26	95	96	−1	+2.9	8.4
24	F	June 26	87	93	−6	−2.1	4.4
25	M	June 27	161	164	−3	+0.9	0.8
26	M	June 27	99	107	−8	−4.1	16.8
27	M	June 27	108	103	+5	+8.9	79.2
28	F	June 28	212	235	−23	−19.1	364.8
29	F	July 2	32	35	−3	+0.9	0.8
30	M	July 2	80	82	−2	+1.9	3.6
31	F	July 2	116	118	−2	+1.9	3.6
32	F	July 2	154	152	+2	+5.9	34.8
33	M	July 3	57	61	−4	−0.1	0
34	M	July 3	82	84	−2	+1.9	3.6
35	F	July 3	76	83	−7	−3.1	9.6
36	F	July 5	84	85	−1	+2.9	8.4
37	F	July 5	183	191	−8	−4.1	16.8
38	M	July 5	227	217	+10	+13.9	193.2
39	F	July 5	75	75	0	+3.9	15.2
40	M	July 8	63	65	−2	+1.9	3.6
			Mean 112.1	Mean 116.0			
			Bias is minus 3.9				

Experimenter: *J. Smith, M. T. (ASCP)*
Name of Reference Method: *Glucose-SN*
Name of Method under study: *Glucose-CD*

made to obtain a new sample from the patient as soon as possible and to investigate any clinical or treatment factor which might account for the discrepancy.

It is evident then that this scheme involves the performance of a minimum of 164 determinations, or more if large discrepancies are encountered.

TABLE 4

LARGE DISCREPANCY FORM

1 No. of Test	2 Sex of Patient	3 Date First Test	4 Result Reference Method	5 Result Test Method	6 Date Second Test	7 Result Reference Method	8 Result Test Method	Clinical Data
1	M	June 20	84	131	June 22	88	82	No obvious reason for the discrepancy on June 20. 34-yr.-old male receiving only salicylates for acute bursitis. No diabetes.

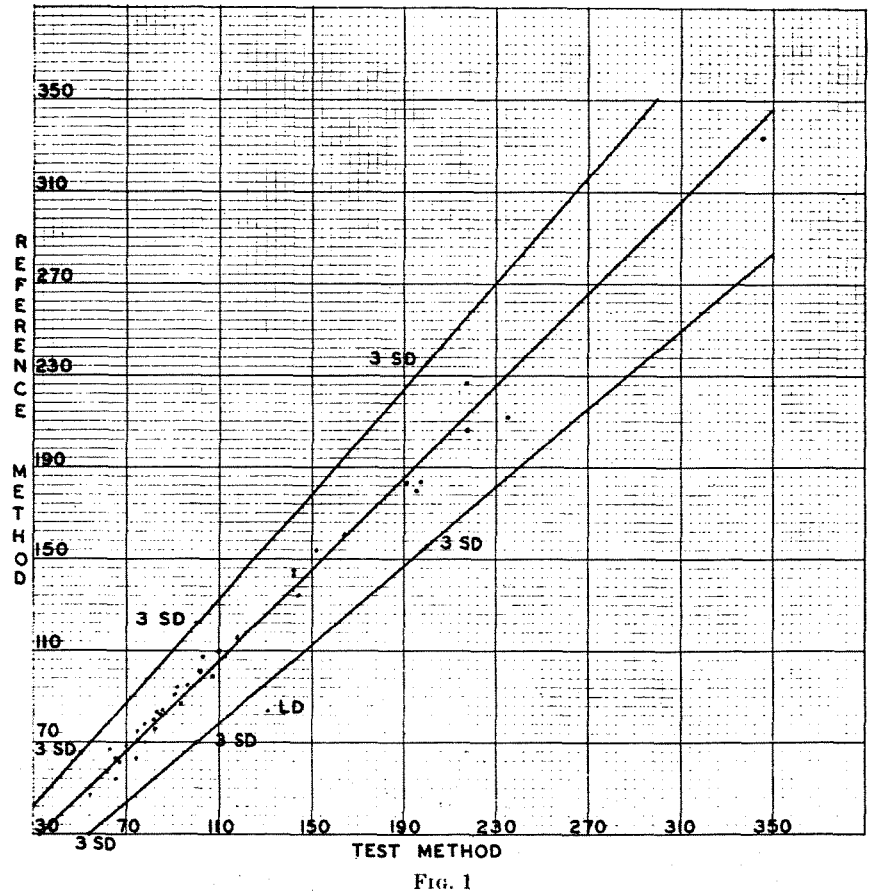

FIG. 1

STATISTICAL MANIPULATIONS

1. Using the data form (Table 3) with the 40 tests run by the 2 methods, proceed as follows:
   A. Determine the arithmetical mean for column 4. Add all figures, divide by number of tests. In the example the result was 112.1.
   B. Similarly determine the arithmetical mean for column 5. In the example the result was 116.0.
   C. Subtract B from A. This indicates the "bias" of the test method and carries a + or − sign. In the example bias is minus 3.9.
   D. On ordinary graph paper (Fig. 1)

mark "reference method" on vertical axis and "test method" on horizontal axis. Mark off appropriate and equal numbers on the 2 axes, with the lowest values in the left lower corner. Plot all of the values from the form using 1 point for each pair of tests. This point is determined by plotting the result of the reference method on the vertical axis and the test method on the horizontal axis.

E. Use the bias figure above (C) and mark it off from the left lower corner. If it is a positive value mark it off on the vertical axis; if it is a negative value mark it off on the horizontal axis. Draw a 45 degree line from this point to the right upper section of the graph.

2. Now refer to the reproducibility forms (Table 1).

A. Begin with the low columns. Take the S.D. of the reference method (in our example found to be S.D. of 5.0) and the S.D. of the test method (4.37). Then apply the formula:

$$\text{S.D. (combined)} = \sqrt{\text{S.D.}^2\text{ (reference)} + \text{S.D.}^2\text{ (test)}}$$

In the example this will be:

$$\text{S.D.} = \sqrt{25 + 19.1} = \sqrt{44.1} = 6.64$$

Mark the point on the line corresponding to the average of the low values of the reference method (use 50 in the example) and plot 3 S.D. (18.9 in the example) above and below this point.

B. Perform a similar calculation for the intermediate column. In the example, use mean of 100 and S.D. of 6.0 for the reference method. This gives a combined S.D. of 8.66. Therefore at the 100 point on the vertical axis mark off ±3 S.D. or 26.

C. Perform a similar calculation for the high column. In the example, use a mean of 200 and S.D. of 8.0 for the reference method. This gives a combined S.D. of ±13.75. Therefore at 200 point on the vertical axis mark off ±39.8.

D. Connect the 3 S.D. above points by a line and the 3 S.D. below points by a line.

E. Any points already plotted (1 D) which fall outside these lines on the graph may be regarded as large discrepancies and should be omitted from the following calculations. They should be added to the large discrepancy form. In the example the only point outside the 3 S.D. lines was already noted on the large discrepancy form. Plot any points from the large discrepancy form on the graph in a different manner and do not use in subsequent calculations.

3. Perform Student's "t" test as follows:

A. Using bias figure from 1 C above fill in columns 7 and 8 of the data form (Table 3). Do not forget that a positive bias should be *subtracted* from, and a negative bias *added* to the column 6 values to yield the column 7 values.

B. Add the values in column 8.

C. Calculate S.D. of the difference by the following formula:

$$\text{S.D.} = \sqrt{\frac{\text{Sum of differences}^2\text{(B)}}{N - 1}}$$

In the example:

$$\text{S.D.} = \sqrt{\frac{1500.4}{39}} = 6.2$$

D. $$t = \frac{\text{bias (1(C))} \times \sqrt{N}}{\text{S.D.}}$$

In the example:

$$t = \frac{3.9 \times \sqrt{40}}{6.2} = \frac{3.9 \times 6.32}{6.2} = 3.97$$

E. A value greater than 2.0 is evidence for a bias.

4. Make additional check for bias as follows. In column 7 count the + and − signs. The minimal permissible number of each is noted in the following table, when N is the number of tests and M the corresponding minimal number of plus or minus signs.

N	50	48	46	44	42	40	38	36
M	17	16	15	15	14	13	12	11

In our example of 40 pairs there are 22 +

and 18 − numbers, each within acceptable limits.

5. Test for "runs." A run is a series of numbers having the same sign. For example, the series −5, −4, 2, 4, −3, −6 consists of 3 runs: −5, −4; +2, +4; and −3, −6. When a 0 value occurs flip a coin to designate a plus or minus sign.
   A. Check column 7 for runs. In the example there are 28.
   B. Arrange values in column 4 in ascending order. List each corresponding column 7 value in a second column. Count the runs in this second column. In the example there are 18.
   C. The minimal number of runs permissible for the number of tests is as follows, when N is the number of tests and M the corresponding number of runs.

N	50	48	46	44	42	40	38	36
M	19	18	17	16	15	14	13	12

   In the example there are more than the minimal number of runs when the numbers are arranged as in A and B above.
6. Test for trends with time as follows:
   A. Take differences from column 6. The first number is listed as X_1, the second X_2, etc.
   B. Make columns of

$$X_1 - X_2 = d_1 \qquad d_1^2$$
$$X_2 - X_3 = d_2 \qquad d_2^2$$

   etc.
   C. Take sum of the d^2 column. Divide by S.D.^{2} (using S.D. from 3 C above). In the example the sum of $d^2 = 3783$, divided by 6.2^2 or 38.4 is 98.5. The ratio so derived should exceed 1.6. A ratio lower than 1.6 is evidence of a trend with time. In the example the value of 98.5 is evidence against such a trend.
7. Further tests for internal discrepancies:
   A. Take the 40 (more or less) figures from column 7 and divide into 2 groups as follows:
      (1) Men *versus* women.
      (2) First 20 *versus* second 20.
      (If there are more or less than 40 divide total number by 2. If an odd number results divide it into 2 parts as nearly equal as possible; for example, 19 and 18).
      (3) Odds *versus* evens. (Nos. 1, 3, 5, etc. *versus* Nos. 2, 4, 6, etc.)
   B. For each pair of these contrasted groups calculate the mean (keeping the signs correct, add the figures and divide by the number of figures so totaled).
   C. Using the formula in 3 C, also calculate the S.D. for each group.
   D. For the 2 members of the groups contrasted the mean and S.D. should be about the same.

   The results of the calculations in the example are as follows:

	Mean of Difference	S.D. of Difference
Men	−0.16	6.0
Women	−0.19	6.1
1st 20	−0.55	5.6
2nd 20	+0.70	6.4
Odds	−0.10	5.0
Evens	+0.25	7.0

   In no instance is there a significant difference between the contrasted groups.

INTERPRETATION OF VALUES

1. Compare the reproducibility forms (Table 1) for the 2 methods.
   A. *Means.* These serve as an index of *accuracy.* The closer to the true value the more accurate is the method.
   B. *Standard deviations.* These serve as indices of *precision.* A difference of less than 30 per cent between the S.D. of the 2 methods is not significant. The smaller the S.D. the more precise is the method.
   C. *Range.* This is another way of expressing precision. The smaller the range the more precise is the method.
2. *Recovery experiments* (Table 2).
   A. Note the difference between observed recovery (column 2) and the expected recovery (column 3) at each level.
   B. Note the expected S.D. for each level in column 1 This can be done from

either the reproducibility form (Table 1) or from the graph (Fig. 1).

C. The difference (A above) should not exceed 2 S.D. (B above) at any level. In the example our results may be tabulated as follows:

Recovery No.	Difference from Expected	± 2 S.D. Limits
B	2 mg.	11.7 mg.
C	1 mg.	14.0 mg.
D	2 mg.	17.2 mg.

Therefore, all of the recoveries are well within the allowable limits.

3. *Large discrepancies* are those which fall outside the 3 S.D. lines of the graph, whether originally listed on the large discrepancy form or not. It is possible for values originally regarded as large discrepancies and so charted actually to fall within the lines. They are then no longer regarded as large discrepancies, but it is not necessary to go back and use them in the data form calculations.

There should not be more than 3 large discrepancies for the 40 samples compared. If there are more than 3 an explanation should be sought, by observing the graph and by study of the clinical data. If all occurred in a very abnormal high range the problem might not be of clinical significance. Or if all were associated with ingestion of a specific medication, allowance might be made for this factor. If, however, these large discrepancies are numerous and random, one can conclude that the test method will not yield results consistent with those of the reference method. It should be mentioned that in some of our studies it was the reference method which gave incorrect values as determined by collateral data.

4. *Bias* (1 C). Ideally this should be small or absent. If a bias appears ("Statistical Manipulations," 1 C) one must question whether it has been confirmed statistically. A "t" test value of more than 2.0 indicates that the bias really exists, as in the example where a value of 3.97 was obtained. If the "t" test value is less than 2.0, one may conclude that a bias has not been confirmed, although it might be confirmed if the series were larger. If no bias is found, or if it is very small as determined by its relation to normal clinical values, it may be ignored. Even if it is large it would not necessarily preclude the use of the test method because new normals could be established and the personnel who use the data could be so instructed. This is well illustrated in the change from "total reducing substances" to so-called "true blood sugar" methods, which has been successfully performed in many clinical laboratories, although a bias of about 20 mg. per 100 ml. existed.

5. *Other tests* (sections 4, 5, 6, and 7 of "Statistical Manipulations"). These are all planned to substantiate the random design of the experiment. If any of the results deviate from the allowable values an explanation must be found. The responsible factor must be identified, explained, and if necessary, corrected before the data can be used. For example, a trend with time (section 6) might be due to progressive spoilage of reagents, and all the data would be suspect.

DISCUSSION

The methods for statistical analysis presented here are explained in many statistical works.[1, 2] Their utilization in the form outlined in this scheme is a practical application not familiar to many chemists and other analysts. Even after the recommended collections of data, statistical manipulations, and interpretations there is still a necessity for value judgments. For example, if the S.D. of the test method is 4 units and of the reference method 2 units, is this significant in the practical application of the data? Only the user can decide this. If the test method is less precise but much quicker and cheaper, and there is no need for extreme precision, one might choose it over the reference method. The decision can be based on the clear definition of the precision, accuracy, and random variability of both methods given by the scheme.

If the original author of a method presents the data of the complete scheme, must each laboratory using the author's method repeat all 164 determinations? This is not necessary. A satisfactory short-cut follows:

1. Perform 40 comparison tests against your present method, as outlined in the "Collection of Data" section.
2. Graph the results. Use your own known S.D. for the reference method and the author's S.D. for the test method and combine them as discussed in "Statistical Manipulations." Determine the bias as described. Graph the ± 3 S.D. lines.
3. From these data one has determined the bias, and the number of large discrepancies, and can proceed to calculate whether there were any flaws in the internal experimental design. If the data indicate satisfactory performance one can proceed with adoption of the new method. If any discrepancies appear, further tests must be done before the method is put into use.

SUMMARY

A detailed description is given of a statistical scheme by which a test method for quantitative analysis can be compared with a reference method. It involves at least 164 determinations and is recommended for use whenever a new method is described or thoroughly evaluated. A "short-cut" method requiring analysis of 40 routine samples by the test and reference method is suggested for use by routine laboratories considering the adoption of an already well-tested method.

REFERENCES

1. Dixon, W. J., and Massey, F. J.: Introduction to Statistical Analysis, Ed. 2. New York: McGraw-Hill, 1957.
2. Youden, W. J.: Statistical Methods for Chemists. New York: John Wiley and Sons, Inc., 1955.